

Кириленко Світлана Едуардівна,

ORCID: <https://orcid.org/0000-0001-8671-1652>,

молодша наукова співробітниця,

відділ наукового формування національних реферативних ресурсів,

Інститут інформаційних технологій,

Національна бібліотека України імені В. І. Вернадського,

Київ, Україна

e-mail: svkirilenko65@gmail.com

Поліщук Тетяна Валентинівна,

ORCID: <https://orcid.org/0000-0002-5157-8182>,

молодша наукова співробітниця,

відділ наукового формування національних реферативних ресурсів,

Інститут інформаційних технологій,

Національна бібліотека України імені В. І. Вернадського,

Київ, Україна

e-mail: polishchuktetana@gmail.com

РБД «Україніка наукова» як джерельна база для створення великої мовної моделі

Проаналізовано переваги впровадження української великої мовної моделі (ВММ) як складової інформаційного пошуку. Розглянуто етапи тренування ВММ та взаємодію з open data. Окреслено адаптацію реферативної бази даних (РБД) «Україніка наукова» до реалій цифрового простору. Розкрито реалізацію на сучасному етапі ВММ «Сяйво» як інструмента трансформації пошуку.

Ключові слова: реферативна база даних, пошуковий запит, велика мовна модель (ВММ) «Сяйво», штучний інтелект.

Сучасна людина не уявляє себе без Google-пошуку зі штучним інтелектом (ШІ). «Запитай у Гугла» – мегапопулярна опція, затребувана всіма верствами

населення будь-якої країни нашої планети. Природно для людини вести пошук через текст, тому підрозділ компанії Google Deep Mind, який займається дослідженнями та розробкою ШІ, створив мовну модель Gemma, яка презентує сімейство відкритих мовних моделей (open language model). Мовна модель апріорі працює з текстом, розуміє й генерує текст зрозумілою мовою, відповідає на запит, може типізувати документи, перекладати між мовами, правильно інтерпретує фразеологізми чи меми. Gemma – не комерційна модель, а створена відкрито та доступно для будь-якої країни світу задля місцевого застосування й навчання на відкритих даних (open data) фахівцями data science, які аналізують величезні масиви інформації. Gemma функціонує на базі англійської мови, виробник намагається привнести її головні принципи в інші мови, але всі вони мають структуру англومовного перекладу, тому кожній окремій державі необхідно створювати власну національну велику мовну модель – Large Language Model (LLM) (системи ШІ, що стане базою майбутніх цифрових сервісів). Процес адаптації LLM під будь-яку національну мову (наприклад, польську, українську або іншу мову) відбувається багатоетапно: «донавчання» (fine-tuning), інструкційне навчання (instruction tuning) та процес вивіряння / вирівнювання (alignment). Етап «донавчання» є процесом додаткового тренування базової Gemma на масштабному корпусі текстів національною мовою (книги, бази даних (БД), урядові документи, наукові джерела тощо), починаючи з «преднавчання» (pre-training). Результатом є удосконалене знання LLM національної мови, стилю та розуміння національного контексту, стійких фразеологізмів, мемів тощо. На інструкційному етапі за допомогою методів контрольованого навчання (supervised fine-tuning) LLM «донавчають» здійснювати операції у вигляді діалогів «запит – відповідь» національною мовою. Наприклад, відповідати на запитання, перекладати, тезово формувати відповідь чи масив тексту, підтримувати бесіду, припустимо, українською мовою. Процес вирівнювання (alignment) важливий для якості національної LLM, на цьому етапі модель налаштовують коректною реакцією на будь-які види запитів, а саме складні, вразливі чи маніпуляційні, тобто визначають поведінку моделі за дійсних умов

вживання, наскільки нешкідливими, придатними будуть її відповіді.

У 2025 році Україна приєдналась до кола країн, які розвивають національну велику мовну модель (ВММ), а саме Міністерство цифрової трансформації України оголосило про створення власної національної ВММ. Всенародним опитуванням через Дію українському ІІІ обрали ім'я «Сяйво». Автори вважають, що назва «Питай», яка теж пропонувалась до обрання, більш відповідає сенсу проєкту створення української мовної моделі, але опитування було не фахове, а публічне, мало хто заглиблювався в тему і перемогла назва перша у списку. Як усі національні ВММ, українська тренується на базі сімейства моделей Gemma від Google – стратегічного партнера проєкту, зокрема Gemma 3 та Gemma 4. Ці моделі працюють з технологією RAG (англ. *Retrieval-Augmented Generation* – генерація з доповненням пошуком), яка поєднує ВММ із зовнішніми БД і дозволяє не вміщати увесь пакет знань до моделі, а підключати її до місцевих баз знань. Далі ВММ шукає затребувану інформацію в національних джерелах і використовує її під час відповіді.

Для високої верифікації інформації відкритих загальних джерел інформації недостатньо, тому фахівці data science проєкту ВММ «Сяйво» співпрацюють з державними органами, медіа, університетами, бібліотеками та іншими інституціями, які опрацьовують терабайти унікальних достовірних українських даних. Національна бібліотека України імені В. І. Вернадського приєдналась до проєкту ВММ «Сяйво», а саме репрезентує масиви реферативної бази даних (РБД) «Україніка наукова», головною перевагою якої є відкрита і загальнодоступна інформація з 28 наукових напрямів (від гуманітарних до медицини, точних і природничих). Наукові масиви РБД «Україніка наукова» поповнюються від 1992 року і донині, цей процес не зупиняється. Якісний науковий контент, авторські першоджерела (а не схоластичні ремарки, що це важливо, бо це важливо), наукові терміни та звирені формули, які вносяться в форматі html з усіма необхідними верхніми/нижніми індексами, математичними функціями (косинус, синус, інтеграл та ін.), грамотно відображаються індекси в хімічних сполуках, згідно з валентністю (без спрощеного H₂O). Основні типи доку-

ментів, що презентовані реферативною інформацією в БД – це монографії, автореферати дисертацій, статті з наукової фахової періодики, підручники ЗВО тощо. Великим плюсом є досвідчені науковці, які професійно компонують тексти саме в логіці будови української мови, що є фундаментальним для навчання української ВММ. Тоді «розмови з ШІ» не будуть проходити через подвійний машинний переклад, коли частково втрачається логіка і контекст відповіді на запит, і результат пошуку буде більш релевантним до запиту. Враховуючи вимоги цифрового суспільства до міжнародної наукової інформації, а саме гнучкість, актуальність представленої інформації, її генерація та інтеграція у світовий науковий контент, РБД «Україніка наукова» забезпечує ефективне функціонування, вільний доступ до результатів наукової діяльності вітчизняних та світових науковців, створює актуальний зріз поточних тенденцій розвитку української та світової науки для фахової орієнтації у величезному масиві інформації.

У межах співпраці з проєктом ВММ «Сяйво» НБУВ гарантує доступ до структурованих масивів різнотематичних текстів РБД «Україніка наукова» (понад 862 тис. рефератів наукових публікацій на червень 2026 р.). На даний момент процес тренування ВММ на завершенні етапу вивірювання (alignment), наближається до тестового періоду. Завдячуючи значному датасету РБД «Україніка наукова» з релевантним матеріалом для навчання мовної моделі й цінному текстовому масивові для машинної обробки актуальної наукової термінології, процес beta-тестування української ВММ очікується прогнозовано вдалим. Зазначимо, що Міністерство цифрової трансформації керуватиме технологічною складовою проєкту, а саме перебігом тренування моделі, залученням експертів та взаємодією з ІТ-індустрією.

Svitlana Kyrylenko,

ORCID: <https://orcid.org/0000-0001-8671-1652>,

Junior Researcher,

Department of Scientific Formation of National Reference Resources,

Institute of Information Technologies,
V. I. Vernadskyi National Library of Ukraine,
Kyiv, Ukraine

e-mail: svkirilenko65@gmail.com

Tetyana Polishchuk,

ORCID: <https://orcid.org/0000-0002-5157-8182>,

Junior Researcher,

Department of Scientific Formation of National Reference Resources,

Institute of Information Technologies,

V. I. Vernadskyi National Library of Ukraine,

Kyiv, Ukraine

e-mail: polishchuktetana@gmail.com

Abstract database «Ukrainika naukova» as a source base for creating a large language model

The advantages of implementing the Ukrainian large language model (LLM) as a component during of information search are analyzed. The stages of LLM training and interaction with open data are considered. The adaptation of the abstract database (RBD) «Ukrainika naukova» to the realities of the digital space is outlined. The implementation of the LLM «Syaivo» as a search transformation tool at the current stage is disclosed.

Keywords: abstract database, search query, large language model (LLM) «Syaivo», artificial intelligence.